Tagging Talkers in a Noisy, Reverberant Large-aperture Microphone-Array Environment

Brian Reggiannini[1], Xiaoxue Li[2], Harvey F. Silverman[3]

Laboratory for Engineering Man/Machine Systems (LEMS), School of Engineering, Brown University
Providence, RI 02912 USA

[1]brian.reggiannini@gmail.com; [2]xiaoxue_li@brown.edu; [3]hfs@lems.brown.edu

Abstract

The problem of associating source tags for conversational speech segments from an unknown number of initially unenrolled talkers who are moving about the focal volume of a large-aperture microphone array is addressed. The problem generally falls under the rubric of speaker diarization, but its conditions and requirements largely separate it from other work. We want to tag in real time on the order of five to ten moving talkers in a reverberant, noisy room from short segments of conversational audio with widely-varying noise conditions. Given that an array system usually locates sources, spatial information may be combined with diarization output to develop the tags. Unfortunately, spatial information alone is inadequate when talkers move while silent, cross positions, etc. Remote microphones and varying-noise conditions make it difficult to collect sufficient training data for static models. Thus our algorithm is trained off-line using only noise-free data and runs in real time. This is possible because a set of robust features that represent the differences between a pair of speech segments is used. The output of the algorithm is a probability, rather than a measure requiring threshold tuning, which makes on-the-fly decisions straightforward. While quite simple, the algorithm has been shown to outperform some commonly-used talker-distance metrics for real data taken from a noisy-environment, large array system. This microphone-array dataset is available to other researchers at
www.lems.brown.edu/array/data/movingtalkers/.

Keywords

Online Diarization; Microphone-array; Talker-similarity

Introduction

We address the problem of real-time, moving-talker tagging in a noisy, reverberant, large-aperture, microphone-array environment using no *a priori* knowledge of any particular talker. This problem is closest in category to speaker diarization as both involve answering the question "which talker spoke when?" for a conversation among several people [1], [2]. For our problem, a small number of talkers (10 or fewer) converse while moving freely in a reverberant room surrounded by a large-aperture microphone array. We observe in this conversational speech that talkers take very short turns (1-20 seconds). Our goal is to produce talker-identity tags continuously with as little latency as possible. We want to process data in a real-time, causal manner given the number of and identities of the talkers are unknown *a priori*. Our problem is made very difficult in that remote microphones in a modest room have an SNR that is about 40dB lower than would a close-talking microphone. Thus each microphone signal exhibits widely varying, high-amplitude "noise" from changing directionality of the talker, changing reverberation patterns, and any standing-wave patterns from periodic background noise such as fans etc. However we do have multiple microphone signals as well as the ability to use spatial source-location data that has been gathered by separate algorithms. For this paper, we assume that the conversation consists of a sequence of *speech segments* each from an individual talker. To the best of our knowledge, this real-world problem has not been addressed for such a harsh environment.

To be real-time and low latency, we needed a method that is sufficient for the specific task and that also requires only modest training and computation. The method discussed here is extremely simple, but it provides an effective way of discriminating between talkers under difficult real-world conditions.

Our method starts off with no individual speaker models and runs on the sequential speech segments, building as the conversation ensues. The method is based upon the formation of a single feature vector concatenated from the robust features of a pair of speech segments, either from the same speaker or from different speakers. Two models, one for "same" speakers and one for "different" speakers, using these concatenated feature vectors, based on clean speech only, are derived, by the EM algorithm from all the training data. When test data is to be tagged, we form a feature vector from the new data and data from a previously-obtained segment and use the models to calculate the probabilities that the feature vector is from the same speaker. Then, decisions may be made.

The contributions of this method are:

- We model pairs of speech segments in a single feature vector, rather than model individual talkers from sequences of feature vectors. This allows us to take advantage of the relative differences between two speech segments, making our features more robust.
- We use robust features. This allows speech models to be trained from a readily-available, clean-speech corpus and still be very useful when applied directly to our noisy data. No development dataset is needed to tune thresholds to the target environment.
- We define a probabilistic talker-similarity metric so there is no need for explicit, feature-dependent thresholds. We explicitly model the between-talker and within-talker variability of our speech features with respect to a large talker population.

Speaker diarization would be difficult when mismatched training and testing conditions exist. Our room environment is extremely noisy, with a significant source of fan noise and a T_{60} of approximately 550 ms [3], [4]. Reverberant noise can cause dramatic differences for various source positions and orientations within the room. For this reason, it would be difficult to acquire sufficient data for training over all conditions. Therefore, we decided to focus on methods that allowed us to train models using only clean, offline speech.

We can view our talker-tagging problem as a series of speaker-verification problems. That is, given two disjoint speech segments, we wish to estimate the probability that both segments were produced by the same talker. Because of high noise levels in our test environment, we expect that, in many cases, we will not be able to make confident decisions from a short length of speech information alone. We have chosen to output a probability so that other sources of information (i.e., talker position) may be combined with speech information in a straightforward way. Also, depending on the target application, a speaker-clustering algorithm may allow for identity decisions to be deferred until more data is available. Moreover, having a probability score, rather than a strict decision, allows us to easily develop more flexible clustering algorithms.

Our main goal is still the diarization of conversational speech, which will typically consist of relatively short talker turns. As a point, most state-of-the-art speaker-verification comparisons are evaluated with no fewer than 10 seconds of enrolment data and 10 seconds of test data [5]. We require lower latencies for a real-time system. In [6] a state-of-the-art speaker-verification technique has been applied to a stream-based speaker segmentation method. This work cited the difficulty in estimating traditional talker models based on Gaussian Mixture Models (GMMs) from the short speech segments often encountered in conversational speech. They use eigenvoice modelling techniques [7], which allow for talker models to be derived from a sliding window of speech. The sliding window they use, however, is 60-seconds long, and is therefore too large for a real-time system. Our method is different because, instead of modelling individual talkers, we model pairs of speech segments. We essentially model the within-talker and between-talker variation of our features, averaged over many talkers in an off-line database. This technique allows us to take advantage of the specific channel-mismatch conditions of our problem. Because enrolment and test data come from a similar environment, we are able to define a relative distance function that is more robust than commonly-used talker-distance measures. Also, we do not suffer from the session variability problem because our system only operates within a single conversation, i.e. in this paper at least, we keep no memory of past conversations.

In this paper, we present a simple, yet elegant, approach to measuring talker similarity for real-time use. We have not yet incorporated this measure into a talker-clustering algorithm. Real-time clustering algorithms have to make many application-specific decisions and also have the opportunity to correct early mistakes. This makes it more difficult to analyse performance with standard error metrics, such as the diarization error rate (DER) [8], which are more informative for the off-line diarization problem. Therefore, we also introduce a new analysis technique that shows the effectiveness of our method in a more direct manner. We compare our method to the generalized likelihood ratio (GLR) and the Bayesian information criterion (BIC) (see Section V).

The paper is organized as follows: Section II provides an overview of previous work; Section III presents a detailed explanation of the proposed method; Section IV describes the specific features used, and their useful properties; V reports the experimental results.

Previous Work

Much research has been conducted and reported to combat channel-mismatch effects for spectral parameters using various types of beamformers [9], but beamformers can create complex channel effects on their own [10]. Various feature-compensation and feature-warping techniques have also been presented to reduce these effects [11]. Our approach is to develop a specific set of speech features that is more robust to channel effects and does not require compensation. This robustness is possible because we are dealing with conversations among only a handful of people, hence we are willing to sacrifice some speaker-discrimination power. We found pitch to be very insensitive to the noise in our environment while a specific set of spectral features has also been found robust under mismatched training and testing conditions.

Microphone-array-based speaker diarization methods can and do incorporate source-position information. Typically, these methods either require individual talkers to be stationary [12], [13] or restricted to specific regions within the room [14]. It is clear that, once we allow talkers to move freely, especially if they move while silent, position information alone is insufficient to label talker identity. While several methods have been presented for off-line speaker diarization in such environments [15], little work has been presented to address the online problem [12].

There are many algorithms that divide the diarization problem into a segmentation component followed by the clustering [1]. However, there are also diarization algorithms that iterate between segmentation and clustering, using an integrated scheme. Some researchers tried to employ hidden Markov models (HMMs) with Viterbi decoding to capture repeated returns of speakers, adding an extra state for the detection of a new speaker [16]–[18]. The main disadvantages of these is the complexity of their models and the unavailability for online applications, as the entire stream is needed to create the HMM states. Current online systems usually learn the cluster for each speaker online [19], [20] and some even combine visual and audio components together [20], [21].

Our concern here is for real-time, online speaker-diarization systems. These require a mechanism for making on-the-fly specific talker tag decisions from short speech segments. Most compare non-overlapping segments of speech using some distance/similarity metric [1]. There are two main metrics, one is the BIC technique [22], [23] and the other is based on the Gaussian divergence [24], [25]. Typical speaker clustering techniques create individual clusters for each speaker using a hierarchical agglomerative clustering scheme [1], [23]. GMMs are often used to model each cluster [26], [27], and the distance metric between clusters is usually the GLR [1], [23], [28]. Given a pair of speech segments, this measure individually models each segment with a Gaussian distribution and computes the likelihood of the data under this 2-Gaussian model. It then compares this value to the likelihood of the combined data when modeled by a single Gaussian. Typically, thresholds on the GLR are determined empirically so it can be incorporated into an online clustering algorithm [28]. While the BIC has often been used as a distance measure for speaker diarization, it is essentially the GLR with an added penalty factor that depends on the number of observations in a window of speech. The short segment size and the analysis technique used in this paper cause this penalty factor to have little effect (see Section IV), making the GLR and the BIC nearly identical. For simplicity, we focus on the GLR for subsequent discussions.

We have discovered two main reasons why measures such as the GLR are undesirable for our problem. First, the use of empirical thresholds limits the robustness of the system. Surely these thresholds will be very dependent on the noise conditions. For calibration in our noisy room, it would be necessary to collect a large development dataset with many talkers speaking in many different positions and orientations. A related issue is that combining speech and position information under the traditional GLR methodology is not straightforward. While, using the technique discussed in this paper, it is possible to normalize GLR values to produce a probability score, this again would require us to record a development dataset in the target environment. The second shortcoming of the GLR is that it inadequately handles the short speech segments we expect [29]. The GLR compares two speech segments directly to each other and lacks knowledge of the inherent within-talker variability of the speech features being used. We will provide evidence to support these observations in Section IV.

Proposed Method

In diarization systems that make decisions in real time – the phase called sequential *speaker clustering* in [1] – a speech segment is evaluated by comparing its features to that of a talker model, derived from some talker-dependent enrolment speech (and possibly some talker-independent training speech) [30]. Such systems typically reduce the feature space to the output of a one-dimensional functional, a talker-distance or talker-similarity value, often a likelihood-ratio, and decisions are made using a threshold. In our method, we also consider the input to the system to be a pair of speech segments, the test segment and an enrolment segment, and seek whether the speech segments originated from the same talker or from different talkers. However, our method recognizes up front that we would like to produce a probability. Therefore, instead of projecting from feature space to a one-dimensional functional space, we estimate multi-dimensional PDFs directly in a double-dimension feature space (i.e., both feature vectors) using standard training procedures. This implementation is made more feasible for our low-dimension feature set that the targeted problem allows. This simple modification allows us to significantly reduce the complexity of the problem and provides an opportunity to obtain a probability of "same-talker" and thus increases robustness.

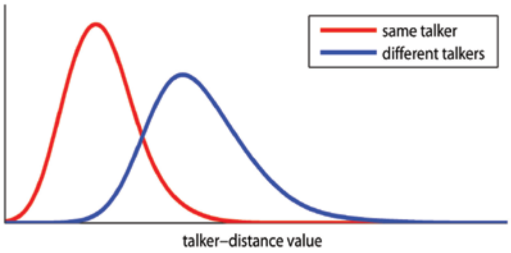

Fig. 1 EXAMPLE PDFS OVER A TALKER-DISTANCE METRIC USED TO ESTIMATE THE PROBABILITY THAT A PAIR OF SPEECH SEGMENTS WERE GENERATED BY THE SAME TALKER

We view the speaker-diarization problem for our application as a simple 2-category classification problem. For each comparison, we produce a single feature vector $\mathbf{X}$ to represent the pair of speech segments in question. $\mathbf{X}$ belongs either to class ω_s, indicating that $\mathbf{X}$ was formed from speech segments from the same talker, or class ω_d, indicating that $\mathbf{X}$ was formed from speech segments from different talkers. We wish to estimate the posterior probability $P(\omega_s|\mathbf{X})$. The prior probabilities $P(\omega_s)$ and $P(\omega_d)$ are assumed to be known. Baye's Rule gives

$$P(\omega_s|\mathbf{X}) = \frac{P(\omega_s)\,p(\mathbf{X}|\omega_s)}{P(\omega_s)\,p(\mathbf{X}|\omega_s) + P(\omega_d)\,p(\mathbf{X}|\omega_d)} \tag{1}$$

Notice that the estimation of a posterior probability is feasible here because we represent the test case with a single observation vector. Traditional methods collect many feature vectors for a given speech segment, assume statistical independence across feature vectors, and compute some form of likelihood ratio. It is unclear whether independence is a fair assumption with such short speech segments. Also, the individual feature vectors are likely to be less robust to noise than some overall segment statistic. While the proposed method may forfeit some talker-discrimination power by reducing a pair of speech segments to a single observation vector, our expectation is that it will gain robustness in the process.

Model Training

A number of practical issues arise when estimating the densities $P(\mathbf{X}|\omega_s)$ and $P(\mathbf{X}|\omega_d)$. These densities are intended to represent the average behavior of the features $\mathbf{X}$ with respect to the target talker population. It is assumed that the number of talkers in the training set is large enough that the individual talkers in the set do not skew the overall distribution. In this case, we assume the training set provides an accurate approximation of the overall characteristics of the target population.

With this in mind, we further split the classes ω_s and ω_d into subclasses based on gender. Class ω_s is divided into two subclasses: ω_s^m for male talkers, and ω_s^f for female talkers. Similarly, class ω_d is split into four subclasses: ω_d^m, ω_d^f, ω_d^{mf}, and ω_d^{fm}. Class ω_d^{mf} represents the case when talker 1 is male and talker 2 is female. Class ω_d^{fm} represents the reversed case. Equations (2) and (3) can be substituted into Equation (1) to apply these gender-specific densities.

$$P(\omega_s)\,p(\mathbf{X}|\omega_s) = P(\omega_s^m)\,p(\mathbf{X}|\omega_s^m) + P(\omega_s^f)\,p(\mathbf{X}|\omega_s^f) \tag{2}$$

$$\begin{aligned} P(\omega_d)\,p(\mathbf{X}|\omega_d) &= P(\omega_d^m)\,p(\mathbf{X}|\omega_d^m) + P(\omega_d^f)\,p(\mathbf{X}|\omega_d^f) \\ &\quad + P(\omega_d^{mf})\,p(\mathbf{X}|\omega_d^{mf}) + P(\omega_d^{fm})\,p(\mathbf{X}|\omega_d^{fm}) \end{aligned} \tag{3}$$

All prior probabilities are assumed unknown. For later use, let us define the prior-probability vector $\mathbf{P}_0$.

$$\mathbf{P}_0 \equiv \left[P\left(\omega_s^m\right), P\left(\omega_s^f\right), P\left(\omega_d^m\right), P\left(\omega_d^f\right), P\left(\omega_d^{mf}\right), P\left(\omega_d^{fm}\right) \right] \tag{4}$$

This formulation shows how to handle mixed-gender test cases. However, because same-gender tests are more difficult and therefore more interesting, we emphasize male-only ($\mathbf{P}_0 = [0.5,\ 0,\ 0.5,\ 0,\ 0,\ 0]$) tests when presenting our results.

We also take into consideration the non-stationarity of speech, which will likely cause our features **X** to be dependent on the durations of the speech segments they represent. Also note that the two speech segments represented by **X** do not necessarily contain the same amounts of data. We have found it best to define a discrete number of segment sizes, and to estimate a separate density for each segment size and subclass. We assume that the segment size can be determined at runtime so that the appropriate density can be applied.

As will be discussed in the following sections, the feature vectors we use in this study are computed over only the voiced frames in an utterance. We therefore define segment size to be the number of voiced frames detected in a speech segment. This measure is more appropriate for this problem than the segment duration itself because of the mismatched training and testing conditions. Our voicing detection algorithm is designed to produce very few false-positives [31]. Therefore, for a given speech segment, we expect to detect fewer voiced frames in our noisy environment than that in our clean training data. The number of voiced frames detected more accurately reflects the amount of speech used to compute the feature vector **X**. The five segment sizes we use for this method can be seen in Table I. The table also illustrates the fact that the segment durations for a given segment size depend on the noise conditions of the data set. The two test databases, labelled "CLEAN" and "ARRAY" are described in the next section. Throughout this paper, all data is sampled at 20 kHz with a frame size of 51.2ms and a frame advance of 10 ms.

For each subclass ω, we estimate densities $p^{(i,j)}(\mathbf{X}|\omega)$ for each pair of segment sizes (i, j). The training data for each density is sampled from the training set of the TIMIT clean-speech database [32]. Each sample is obtained using the following procedure.

1) Two talkers are selected at random, taking into account the gender-specifications of the subclass ω.
2) The numbers of voiced frames for each speech segment are selected uniformly from the ranges given in Table I for the appropriate segment sizes.
3) Two appropriately-sized non-overlapping speech segments are chosen at random, one for each talker.
4) From the pair of speech segments, a single observation vector X is produced.

TABLE I DISCRETE SET OF SEGMENT SIZES USED IN SPEECH MODELS.DURATION RANGES REPRESENT 10% – 90% INTERVALS COMPUTED OVER ALL SAMPLES IN EACH TEST DATABASE.WHEN MEASURING DURATION, LARGE SILENCES WERE REMOVED BY HAND FOR THE ARRAY DATABASE, AND USING THE PROVIDED TRANSCRIPTIONS FOR THE CLEAN TIMIT DATABASE.

Segment Size Index	Number of Voiced Frames	Duration (sec) (10% - 90%)	
		Clean	Array
1	1 - 25	0.0 - 0.5	0.0 - 1.2
2	26 - 75	0.5 - 1.4	0.8 - 3.6
3	76 -150	1.3 - 2.8	2.5 - 7.5
4	151 - 250	2.6 - 4.7	5.0 - 12.4
5	251 - 400	4.4 - 7.4	8.4 - 19.4

This process is repeated to create the training set. Using these random samples, we estimate the maximum-likelihood full-covariance GMM for $p^{(i,j)}(\mathbf{X}|\omega)$ using the EM algorithm. Because each model represents an average over many talkers, each density requires only a small number of Gaussians. In this paper, each PDF is a GMM with 5 Gaussians.

Model Testing

For testing, we recorded a small talker database in our target environment. 60-second recordings of each of 25 male and 25 female talkers were made using 96 microphones from our large-aperture microphone array [3], [4] (these recordings and the corresponding microphone locations can be found at www.lems.brown.edu/arra/data/movingtalkers/). Talkers were asked to speak at a natural volume while walking freely around the 6.5m x 4.7m focal area of the room. Talkers read from a list of sentences from the TIMIT database. Only a single talker was present for each recording. At 10-ms intervals, four indepen-dent source-location estimates were made, with each estimate coming from a 24-microphone locator positioned in a corner of the room. Location estimates were made using the SRP-PHAT-based method described in [33]. A delay-and-sum beamformer with inverse-distance weights focused at the location estimate with the highest SRP-PHAT value was used. For comparison, we also tested on the clean data in the test set of the TIMIT database. We will refer to the noisy database and clean TIMIT database as "ARRAY" and"CLEAN", respectively.

Test samples were generated using the same sampling method used to generate training samples. Each sample consists of two short, non-overlapping speech segments to be compared as "same talker" or "different talkers". Because talkers in the ARRAY database were moving throughout their respective recordings, the sampling procedure produces speech segments that were likely to originate from different positions within the room and, certainly, widely different noise conditions. This database represents the worst-case-scenario in terms of noise conditions.

Talker-Discriminating Features

In the following sections, we describe the specific speech features being used and analyze their effectiveness in discriminating between talkers from speech segments of various durations.

Pitch Information

Talker pitch has long been viewed as an important parameter to convey identity [34], [35], and it has been the focus of many recent studies on speaker recognition [36]–[39]. These studies typically exploit temporal changes in prosodic features such as pitch. The use of such high-level features generally requires significantly more talker-specific training data [36], [39] than is available for our problem. For our study of short speech segments, we have chosen to study static features averaged over the duration of the segment, building upon our previous work [31].

Fundamental frequency, as an approximation of pitch, is estimated using a modfied version of the cepstrum method [40]. This method, described in more detail in [31], was shown there to be fairly robust to noise. For a given speech segment, the scheme first labels each frame as voiced or unvoiced, and then computes the pitch mean and standard deviation over all voiced frames in the segment. For a pair of speech segments, the 4-dimensional vector $\mathbf{X}_{f0}$ is formed by concatenating mean and standard deviation values from each segment.

Figure 2 shows our results for male-only test conditions ($\mathbf{P}_0 = [0.5, 0, 0.5, 0, 0, 0]$), for both the noisy microphone-array dataset (ARRAY) and the original test set of TIMIT (CLEAN). For each dataset and segment-size pair, we selected a random set of test samples $\mathbf{X}^n_{f0}$ indexed by n ∈ [1, N], using a chosen prior-probability vector $\mathbf{P}_0$. We then evaluated the probability P $P(\omega_s | \mathbf{X}^n_{f0})$ (Equation 1) for each sample. A histogram was formed by placing samples into 20 linearly-spaced bins according to their probabilities $P(\omega_s | \mathbf{X}^n_{f0})$. Let $P_{\text{bin}}[k] \equiv 0.025 + 0.05k$ be the center of each bin of width 0.05, for $k \in [0, 19]$, and let m_k be the number of samples placed into the k^{th} bin. In order to evaluate our performance, we used the true talker-identity labels for the test samples to count the number of samples in each bin from each class. Let s_k and d_k be the numbers of samples in the k^{th} bin from classes ω_s and ω_d, respectively.

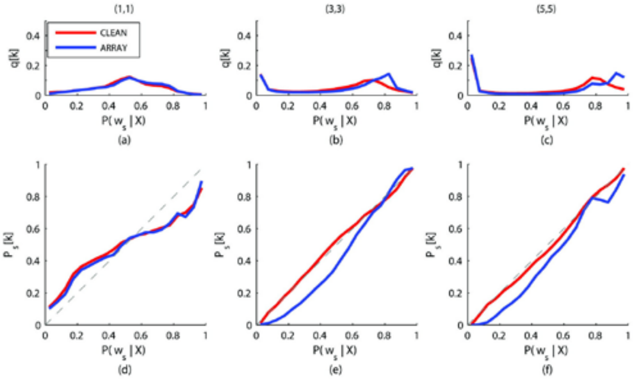

Fig. 2 RESULTS FOR PITCH FEATURES WITH SEGMENT-SIZE INDEX PAIRS (1,1), (3,3), AND (5,5). MALE TALKERS ONLY

Figures 2(a-c) show $q[k] \equiv m_k/N$, the PDF of $P(\omega_s | \mathbf{X}^n_{f0})$. These PDFs tell us about the discrimination power of features $\mathbf{X}_{f0}$. Ideally, all $P(\omega_s | \mathbf{X}^n_{f0})$ values would fall at 0 and 1, for samples of classes ω_s and ω_d, respectively. This would indicate that all decisions could be made with 100% confidence.

Our results show that pitch rarely provided enough information for us to confidently detect an identity match ($P(\omega_s | \mathbf{X}^n_{f0}) \approx 1$). However, for the larger segment sizes, we did find many $P(\omega_s | \mathbf{X}^n_{f0}) \approx 0$, which would allow us to confidently reject an identity match, even though both talkers were male. This trend will clearly be more dramatic for mixed-gender cases.

Next, we used the known talker-identity counts, s_k and d_k, to evaluate the accuracy of our $P(\omega_s | \mathbf{X}^n_{f0})$ measure. We computed the true probability of talkers being the same for bin k, $P_s[k] \equiv s_k / (s_k + d_k) = s_k / m_k$, which is shown along the vertical axes of Figures 2(d-f). We see that $P(\omega_s | \mathbf{X}^n_{f0})$ is highly correlated with $P_s[k]$. This is true even for the ARRAY dataset, indicating that the pitch features $\mathbf{X}_{f0}$ are robust to noise. Note that deviations from the identity line $P_{\text{bin}}[k] = P_s[k]$ are not very significant for probability bins with few observations in them, i.e., $q[k] \approx 0$.

Spectral Information

We also derive spectral features for the voiced frames, as selected by the pitch-detection system, to represent a pair of speech segments. For each voiced frame t, in a pre-emphasized speech segment u, we apply a Hamming window and compute the short-time Fourier spectrum $Y_t^u[r]$, where r is the frequency index. Each spectral vector $Y_t^u[r]$ is normalized by dividing by its maximum magnitude. Let us denote this normalized vector $\hat{Y}_t^u[r]$.

$$\hat{Y}_u^t[r] = \frac{Y_u^t[r]}{\max_r \left| Y_u^t[r] \right|} \tag{5}$$

A log-energy vector $Z_t^u[m]$ is produced by passing the normalized spectral vectors $\hat{Y}_t^u[r]$ through a Mel-scale filter bank $H_m[r]$ [30], where m is the filter index.

$$Z_t^u[m] = \log\left(\sum_{r=r_1}^{r_2} \left| H_m[r] \hat{Y}_u^t[r] \right|^2 \right) \tag{6}$$

The filter bank has 25 overlapping triangular filters on a Mel scale from 500 Hz to 4500 Hz, corresponding to a frequency index range of r_1 to r_2.

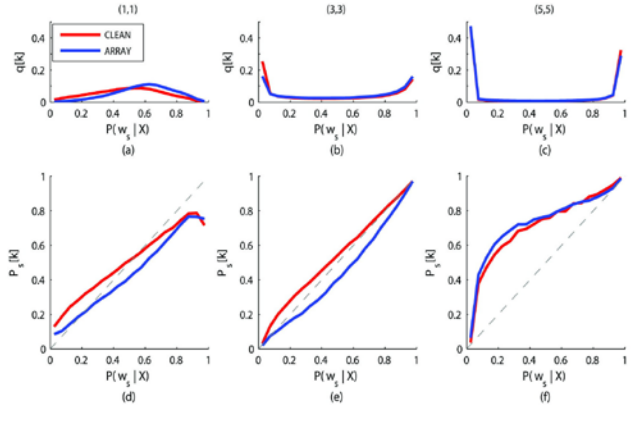

Fig. 3 RESULTS FOR SPECTRAL FEATURES WITH SEGMENT-SIZE INDEX PAIRS (1,1), (3,3), AND (5,5). MALE TALKERS ONLY

For a pair of speech segments u and v, with T_u and T_v voiced frames detected in each, a 25-dimensional feature vector $\mathbf{X}_S$ is created by computing the difference between the means of the log-energy vectors.

$$\mathbf{X}_S^{(u,v)}[m] = \left(\frac{1}{T_u} \sum_{t=1}^{T_u} Z_t^u[m] \right) - \left(\frac{1}{T_v} \sum_{t=1}^{T_v} Z_t^v[m] \right) \tag{7}$$

Because the mean operation will reduce the effects of additive noise, and the difference in the log domain will help combat correlated noise, the vector $\mathbf{X}_S$ is expected to be quite robust. This is a major benefit of the proposed method. Relative-distance features such as these cannot be used under the most other segment-comparison frameworks.

Figure 3 shows our experimental results for spectral features $\mathbf{X}_S$ with male talkers ($\mathbf{P}_0$ = [0.5, 0, 0.5, 0, 0, 0]). From Figures 3(d-f) we see that these spectral features are indeed robust to noise and are well correlated to $P_s[k]$ (for bins with $q[k] >> 0$). From Figures 3(a-c) we see that, given enough data, these features can also discriminate between male talkers. While pitch features $\mathbf{X}_{f0}$ generally only allowed us to reject an identity match with cofidence, the spectral features $\mathbf{X}_S$ also assigned many high probabilities ($P(\omega_s | \mathbf{X}_S) \approx 1$). This indicates that we can very often detect an identity match with high confidence. Our pitch features $\mathbf{X}_{f0}$ and spectral features $\mathbf{X}_S$ thus appear to be complimentary.

Pitch and Spectral Information Combined

Our full system combines both pitch and spectral information into a 29-dimensional feature vector $\mathbf{X}$, formed by concatenating the pitch vector $\mathbf{X}_{f0}$ and the spectral vector $\mathbf{X}_S$. Our results are shown in Figure 4. By comparing Figures 2 and 3 to Figure 4, we see the additional talker discrimination power attained by combining both feature vectors.

We will now define three metrics to help us summarize overall performance. First we want to quantify the error for our $P(\omega_s | \mathbf{X})$ estimates, with respect to the empirical probabilities $P_s[k]$. The error variance σ^2_{err} is defined as

$$\sigma^2_{\text{err}} \equiv \sum_k q[k] \left(P_{bin}[k] - P_s[k] \right)^2 \tag{8}$$

and a lower value of error is preferable.

To assess the ability of the features $\mathbf{X}$ to reject an identity match, we define the mean rejection score μ_{rej}. Let $\{n_0\}$ be the subset of sample indices $n \in [1, N]$ for which $P(\omega_s | \mathbf{X}^n) \leq 0.5$, and let $|n_0|$ be the number of members in the subset.

$$\mu_{rej} \equiv \frac{1}{|n_0|} \sum_{n \in \{n_0\}} P(\omega_s | \mathbf{X}^n) \tag{9}$$

Clearly, as defined, better algorithms will have a lower mean rejection score.

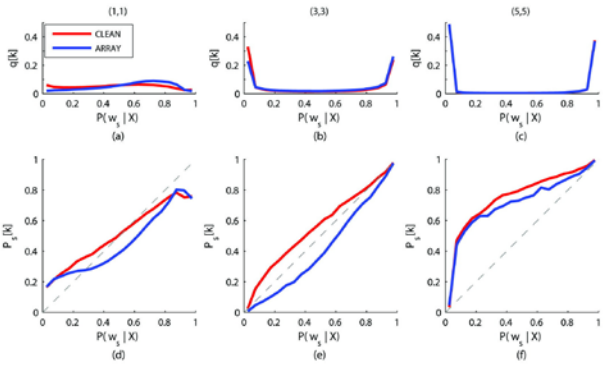

Fig. 4 RESULTS FOR PITCH AND SPECTRAL FEATURES WITH SEGMENT-SIZE INDEX PAIRS (1,1), (3,3), AND (5,5). MALE TALKERS ONLY

Similarly, the mean detection score μ_{det} allows us to measure the effectiveness of a feature set in detecting an identity match within a particular dataset. Let $\{n_1\}$ be the subset of sample indices for which $P(\omega_s|\mathbf{X}^n) \geq 0.5$.

$$\mu_{\det} \equiv \frac{1}{|n_1|} \sum_{n \in \{n_1\}} P\left(\omega_s \,|\, \mathbf{X}^n\right) \qquad (10)$$

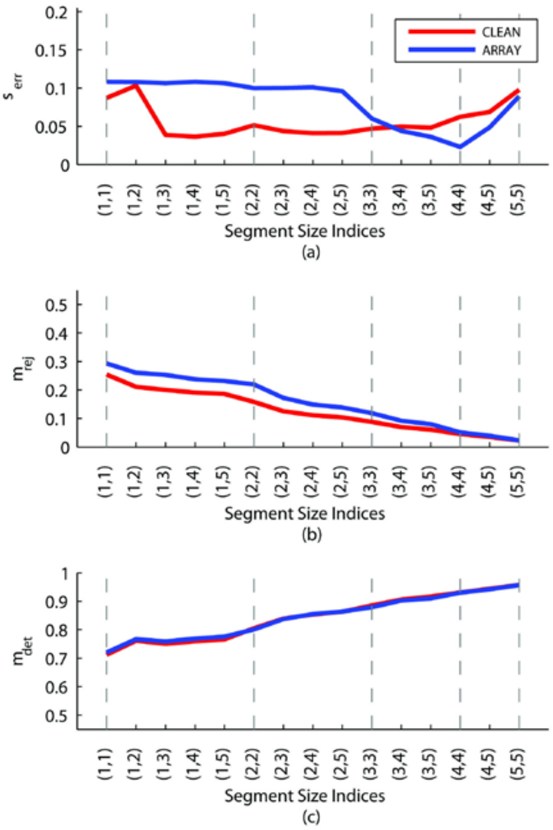

Fig. 5 SUMMARY STATISTICS (A) ΣERR, (B) μREJ, AND (C) μDET, FOR SPECTRAL AND PITCH FEATURES, AND FOR EACH SEGMENT-SIZE PAIR. MALE TALKERS ONLY

For this definition, a higher value of the mean detection score will imply the better algorithm.

These summary statistics are given in Figure 5, where each label along the horizontal axis represents a pair of segment sizes, as defined in Table I. Figure 5(a) shows that we achieve similar amounts of error σ^2_{err}, for both the ARRAY and CLEAN datasets. Figures 5(b,c) shows an increase in discrimination power as the speech segment sizes increase, with μ_{rej} approaching 0 and μ_{det} approaching 1.

Figure 6 shows the summary statistics for three different gender conditions for the ARRAY dataset. In Figure 6(a), we see that the female models appear to introduce more error than the male models. This could be due to the fact that the female training set has significantly fewer talkers than the male training set. Not surprisingly, Figures 6(a,b) show the increase in discrimination power achieved when mixed-gender scenarios are possible. This is especially true of the rejection score μ_{rej}, which shows that we can often discriminate between talkers of different genders with high confidence.

Proposed method vs. Standard measures

The analysis technique presented has allowed us to measure both the robustness of a given feature set and

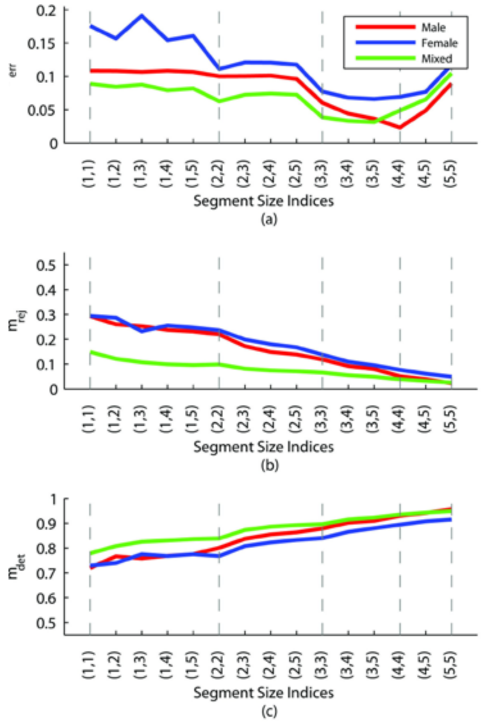

Fig. 6 SUMMARY STATISTICS STATISTICS (A) ΣERR, (B) μREJ, AND (C) μDET, FOR SPECTRAL AND PITCH FEATURES, AND FOR EACH SEGMENT-SIZE PAIR, FOR THE ARRAY DATASET. (MALE) MALE TALKERS ONLY (P_0 = [0.5, 0, 0.5, 0, 0, 0]). (FEMALE) FEMALE TALKERS ONLY (P_0 = [0, 0.5, 0, 0.5, 0, 0]). (MIXED) MALE AND FEMALE TALKERS (P_0 = [0.25, 0.25, 0.125, 0.125, 0.125, 0.125])

the ability of that feature set to discriminate between talkers. For comparison, this same technique can be applied to the commonly-used GLR [1], [15], [20] and BIC [14], [15] measures.

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be the sets of D-dimensional feature vectors obtained from each speech segment in question. Let $N(\mathbf{X}|\mu_1, \Sigma_1)$ be a multivariate Gaussian distribution estimated to fit data $\mathbf{X}_1$. $N(\mathbf{X}|\mu_2, \Sigma_2)$ is similarly estimated from observations $\mathbf{X}_2$. Finally, let $N(\mathbf{X}|\mu_c, \Sigma_c)$ be the Gaussian distribution estimated from set X_c , formed by concatenating $\mathbf{X}_1$ and $\mathbf{X}_2$ into a single set. The GLR is defined as

$$GLR \equiv \frac{L(\mathbf{X}_1 | \mu_1, \Sigma_1) L(\mathbf{X}_2 | \mu_2, \Sigma_2)}{L(\mathbf{X}_c | \mu_c, \Sigma_c)} \tag{11}$$

where $L(\mathbf{X}|\mu, \Sigma)$ represents the joint likelihood of model $N(\mathbf{X}|\mu, \Sigma)$ given feature set $\mathbf{X}$. All observations are considered independent when computing $L(\mathbf{X}|\mu, \Sigma)$.

The BIC, as used for speaker change-detection and clustering, is the same likelihood ratio with a penalty factor P [15].

$$BIC \equiv \frac{L(\mathbf{X}_1 | \mu_1, \Sigma_1) L(\mathbf{X}_2 | \mu_2, \Sigma_2)}{L(\mathbf{X}_c | \mu_c, \Sigma_c)} - \lambda P \tag{12}$$

The constant λ is typically set to 1, and the penalty factor P is defined as,

$$P \equiv \frac{1}{2}\left[D + \frac{1}{2}D(D+1)\right]\log N \tag{13}$$

where N is the number of D-dimensional feature vectors in the combined set X_c.

We repeated the analysis used to create Figure 5 using the GLR measure as a one-dimensional feature vector X_{GLR}. For speech segments u and v, with voiced frames $t \in [1, T_1]$ and $t \in [1, T_2]$ respectively, we concatenated each $Z_t^u[m]$ and $Z_t^v[m]$ with its respective pitch frequency value. Then, we computed the GLR for the pair of speech segments using this combined vector to produce the one-dimensional feature vector X_{GLR}. This is essentially what is done in a typical talker clustering application. A single GLR value is used to measure the distance between a pair of speech segments. All training and testing was carried out as previously described, with separate models used for each discrete segment size defined in Table I. This mitigates the previously-reported problem of GLR values dependent on segment sizes [21]. The same procedure was repeated for the BIC. Our results are shown in Figure 7, with results for the CLEAN database on the left and results for the ARRAY database on the right. In each case, we compare our method (PROPOSED) to the GLR method (GLR) and the BIC method (BIC). Note that while the talker-similarity measures are different, the features extracted from each segment are the same.

TABLE II RATE OF OCCURRENCE OF TEST SAMPLES FOR WHICH THE GLR AND THE BIC COULD NOT BE EVALUATED BECAUSE OF NUMERICAL ISSUES. SHOWN ONLY FOR SEGMENT-SIZE PAIRS FOR WHICH SUCH PROBLEMS

	% Samples Not Evaluated	
Segment Sizes	**CLEAN**	**ARRAY**
(1,1)	87.8	87.7
(1,2)	63.0	62.9
(1,3)	62.7	62.6
(1,4)	62.6	62.5
(1,5)	62.7	62.6
(2,2)	2.0	2.0
(2,3)	1.0	1.0
(2,4)	1.0	1.0
(2,5)	1.0	1.0

Because we group observations according to their segment sizes, the BIC and the GLR are nearly identical. The BIC penalty factor P is approximately the same for all observations with a given pair of segment sizes. For simplicity, we discuss our results with respect to the GLR only.

We now address the previously-described shortcomings of the GLR with respect to our problem. The first issue is that of robustness. The large σ_{err} values shown in Figure 7(b) for the GLR method tell us that any thresholds on GLR determined empirically using the clean training data would not be effective in the noisy environment. Note that while σ_{err} was small for segment sizes (1,1) through (1,5), the GLR offered no discrimi-natory power in these cases ($\mu_{rej} \approx \mu_{det} \approx 0.5$).

Also, the GLR is less informative for short segment sizes. As discussed previously, speech is a non-stationary process. Disjoint speech segments from the same talker can have different statistical properties. The GLR, however, compares the individual segment distributions to each other directly, and will inappropriately assign a large distance in such situations. Also, because the statistical properties of different talkers often overlap, it is quite possible to find two short speech segments from different talkers with nearly identical distributions. The GLR will assign such segments a distance that is too small. The proposed method is more appropriate for these short segments because segment similarity is assessed with respect to a large offline talker database. By taking within-talker and between-talker variability into account for each feature dimension, we were able to achieve better talker discrimination, even with clean data. Figures 7(c-f) support this. Note that as the segment sizes increase, the non-stationarity of speech becomes less of an issue, and the GLR performance approaches that of the proposed method. This did not occur, however, until the largest segment-size pair (5,5). For our noisy environment, this can correspond to speech segments of up to approximately 20 seconds in duration. Because we expect much shorter speech segments in conversational speech, the proposed method is a more appropriate choice.

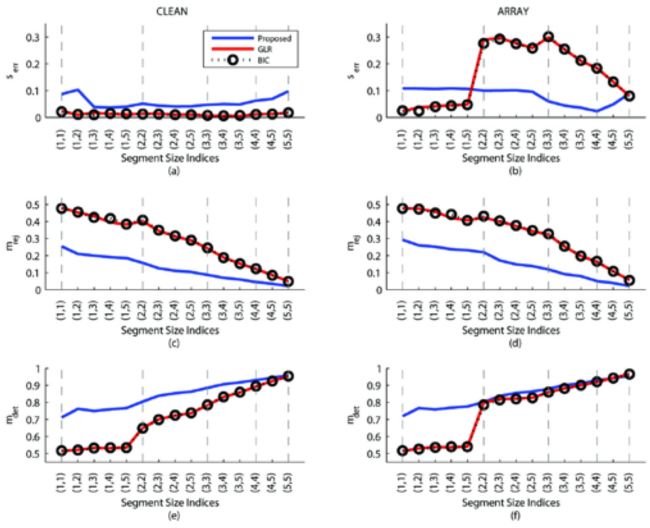

Fig. 7 SUMMARY STATISTICS (A,B) Σ_{ERR}, (C,D) μ_{REJ}, AND (E,F) μ_{DET}, FOR THE PROPOSED METHOD, THE GLR, AND THE BIC. MALE TALKERS ONLY

Computational issues also arise for the GLR and the BIC with short speech segments. Speech segments with few observation vectors can produce singular covariance matrices. In these cases, neither the GLR nor the BIC can be computed. Table II shows the proportion of samples in each dataset for which the GLR and the BIC could not be evaluated. These unevaluated samples were not included in the results shown in Figure 7. The target application would have to decide what to do in such cases. For example, a 0.5 probability could be assigned. The method proposed in this paper did not encounter these kinds of numerical issues.

Conclusions and Future Work

We have introduced a computationally-efficient paradigm for tagging on the order of 10 moving sources in a reverberant, noisy, distant-microphone environment. It can be, and is currently being, implemented for a 128-microphone array [42], in real time on a standard desktop system. Our goal was to develop a system that is simple and **sufficient** to do our task, a method that was trainable using clean data only. As noise conditions vary greatly as a function of talker position and orientation, acquisition of sufficient data for a training database is essentially infeasible. Our technique uses specific robust speech features to overcome, as much as possible, the training/testing mismatch problem. A large part of the robustness comes from the fact that we model the differences between pairs of speech segments, rather than attempting to model individual talkers. We also try to use features that are robust to the kinds of noise that will prevail in the normal-room, remote-microphone situation. The training technique itself nicely produces a probability. Because we compute a probability, rather than make a strict identity decision, our method may be combined with other, more complex probabilistic methods quite easily, if desired. For example, we combine our speech-tagging results with talker-position information in the real-time system under construction.

The proposed method, simply trained on clean data, was shown to outperform the GLR and the BIC for discriminating between talkers from short-speech segments (0-20 seconds), about the lengths of talker segments that occur naturally in conversational speech. The technology can be applied to the real-time transcription of meetings, broadcast news, and courtroom proceedings. Automatic real-time transcriptions could be useful to the deaf and could also be used to improve human-machine interactions.

One limitation of the system in this paper is that it requires only a single talker to be active at a time. The problem of segmenting and dealing with multiple concurrent speakers [43] is currently being addressed as we implement the system in real time. Future work might also involve the design of a clustering algorithm that uses a probabilistic similarity measure like the one suggested here. The proposed method may benefit from the use of a larger training database, one with more speech per talker that would allow an extension for longer speech segments. Finally, one might investigate using a more standard talker-modeling technique once the proposed method has allowed for sufficient enrolment data to be gathered.

REFERENCES

A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP '03*, vol. 4, 2003, pp. 788–791.

A. G. Adami, "Modeling prosodic differences for speaker recognition," *Speech Communication*, vol. 49, no. 4, pp. 277– 291, 2007.

A. M. Noll, "Cepstrum pitch determination," *The Journal of the Acous-tical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

A. Noulas and B. Krose, "On-line multi-modal speaker diarization," in *International Conference on Multi-modal Interfaces, ICMI07*, 2007, pp. 350 – 357.

A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on*, vol. 2, 1998, pp. 757–760.

B. Reggiannini and H. Silverman, "Using mean fundamental frequency for real-time talker labeling in a microphone array environment," 2010, [Online]. http://www.lems.brown.edu/array/download.html.

B. S. Atal, "Automatic speaker recognition based on pitch con-tours," *The Journal of the Acoustical Society of America*, vol. 52, no. 6B, pp. 1687–1697, 1972.

C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *Proc. Fall Rich Transcription Workshop* (RT-04). Palisades, NY, Nov. 2004.

C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Toward robust speaker segmentation: The icsi-sri fall 2004 diarization system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04),* Nov. 2004, pp. 127 – 132.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

D. Liu and F. Kubala, "Online speaker clustering," in *Proc. ICASSP '03*, vol. 1, 6-10 2003, pp. I–572 – I–575.

D. Moraru, S. Meignier, L. Besacier, J. F. Bonastre, and I. Magrin-Chagnolleau, "The elisa consortium approaches in speaker segmentation during the nist 2002 speaker recognition evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, Honk-Kong, China, May. 2003.

D. Reynolds and P. Torres-Carrasquillo, "The mit lincoln laboratory rt-04f diarization systems: Applications to broadcast audio and telephone conversations," in *DARPA EARS RT-04F Workshop*, White Plaine, NY, Nov. 2004.

D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The supersid project: exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP '03*, vol. 4, 6-10 2003.

F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008.

G. Lathoud, I. A. Mccowan, and D. C. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *Proceedings of Eu-rospeech 2003*, September 2003.

H. Do, H. Silverman, and Y. Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *Proc. ICASSP '07*, vol. 1, Apr. 2007, pp. I–121 – I–124.

H. F. Silverman, W. R. P. III, and J. Sachar, "Factors affecting the performance of large-aperture microphone arrays," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2140–2157, 2002.

H. Silverman, I. Patterson, W.R., and J. Sachar, "Early experimental results for a large-aperture microphone-array system," in *Proceedings of the 2000 IEEE Sensor Array and Multichannel Signal Processing Workshop.*, 2000.

I. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proc. 2001: A Speaker Odyssey*, 2001.

J. Ajmera, G. Lathoud, and L. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 1, 17-21 2004, pp. I – 605–8 vol.1.

J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus," *Technical Report NISTIR 4930, National Institute of Standards and Technology*, 1993.

J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in gmm speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proceedings of ICSLP '96*, vol. 3, 1996, pp. 1333–1336.

J. Markel, B. Oshika, and J. Gray, A., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 4, pp. 330 – 337, Aug 1977.

J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1212 –1224, 2007.

J. Stokes, J. Platt, and S. Basu, "Speaker identification using a micro-phone array and a joint hmm with speech spectrum and angle of arrival," in *Multimedia and Expo, 2006 IEEE International Conference on*, 2006, pp. 1381 – 1384.

K. Han, S. Kim, and S. Narayanan, "Robust speaker clustering strategies to data source variation for improved speaker diarization," in *Automatic Speech Recognition Understanding, 2007. ASRU. IEEE Workshop on*, Dec. 2007, pp. 262–267.

K. Ishiguro, T. Yamada, S. Araki, and T. Nakatani, "A probabilistic speaker clustering for doa-based diarization," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, 2009, pp. 241 –244.

K. Markov and S. Nakamura, "Never-ending learning system for on-line speaker diarization," in *Proc. IEEE ASRU07*, Kyoto, Japan, 2007, pp. 699 – 704.

M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97 – 99.

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 19, no. 4, pp. 788 –798, May 2011.

O. Vinyals and G. Friedland, "Towards semantic analysis of conversa-tions: A system for the live identification of speakers in meetings," in *Proc. IEEE ICSC08*, Santa Clara, CA, 2008, pp. 426 – 431.

P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and R. Gadde, "Speaker recognition using prosodic and lexical features," in *IEEE Workshop on Automatic Speech Recogni-tion and Understanding '03*, 30 2003, pp. 19 – 24.

S. Khanal, H. F. Silverman and R. R. Shakya, "A free-source method (FrSM) for accurately and quickly calibrating a large-aperture microphone array". Submitted for publication, currently available at http://www.lems.brown.edu/array/download.html.

S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin, "Evolutive HMM for multispeaker tracking system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, Istanbul, Turkey, vol. II, Jun. 2000, pp. 1201 – 1204.

S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Crete, Greece, Jun. 2001, pp. 175 – 180.

S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

S. Tranter and D. Reynolds, "An overview of automatic speaker diariza-tion systems," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 5, pp. 1557 –1565, Sept. 2006.

The 2009 (rt-09) rich transcription meeting recognition evaluation plan. [Online]. http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf.

T. Koshinaka, K. Nagatomo, and K. Shinoda, "Online speaker clustering using incremental learning of an ergodic hidden markov model," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on,* 2009, pp. 4093 –4096.

Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 15, no. 7, pp. 2023–2032, Sept. 2007.

Brian Reggiannini was born in Boston, MA, on September 15, 1984. He received the Sc.B., Sc.M., and Ph.D. degrees in computer engineering from Brown University, Providence, RI, in 2007, 2009, and 2012, repectively. His graduate research focused on speaker recognition in adverse acoustical environments using microphone arrays.

Dr. Reggiannini is currently working on quadrature error correction algorithms for wireless tranceivers at Analog Devices, Inc. in Norwood, MA. He is a member of IEEE.

Xiaoxue Li was born in Yichang, Hubei Province, P.R.China on December 14, 1985. She received ScB. and Sc.M. degrees in information science and electronic engineering from Zhejiang Unversity, Hangzhou, Zhejiang Province, P.R.China in 2007 and 2010 respectively. She received Sc.M. degree in computer engineering from Brown University, Providence, RI, in 2012.

She is currently a Ph.D candidate in computer engineering of Brown University. Her interest is mainly on speech processing, microphone-array signal processing and pattern recognition.

Harvey F. Silverman was born in Hartford, Connecticut on August 15, 1943. He received the BS and BSEE degrees from Trinity College in Hartford in 1965 and 1966, and the ScM and PhD degrees from Brown University, Providence, RI in 1968 and 1971 respectively, all degrees in electrical engineering.

He worked with H. Joseph Gerber of Gerber Scientific Instruments from 1964-66 and helped design the first Gerber plotter. He was at the IBM Thomas J. Watson Research Center from 1970 to 1980, working in the areas of digital image processing, computer performance analysis, and was an original member of the IBM Research speech recognition group that started in 1972. He was manager of the Speech Terminal project from 1976 until 1980. At IBM he received several outstanding innovation awards and patent awards. In 1980, he was appointed Professor of Engineering at Brown University in Providence, RI, and charged with the development of a program in computer engineering. His research interests currently include microphone-array research, array signal processing, and embedded systems. He has been the Director of the Laboratory for Engineering Man/Machine Systems in the Division of Engineering at Brown since its founding in 1981. From July 1991 to June 1998 he was the Dean of Engineering at Brown University.

Dr. Silverman was a member of the IEEE Acoustics, Speech and Signal Processing Technical Committee on Digital Signal Processing and was its Chairman from 1979 until 1983. He was the General Chairman of the 1977 ICASSP in Hartford, CT. He received an IEEE Centennial Medal in 1984. He was Trustee of Trinity College in Hartford, CT 1994-2003, and is a Lifetime Fellow of IEEE.